ARL-TR-8064 ● July 2017

US Army Research Laboratory

Verification of Weather Running Estimate–Nowcast (WRE–N) Forecasts Using a Spatial–Categorical Method

by John W Raby

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-8064 ● JULY 2017

US Army Research Laboratory

Verification of Weather Running Estimate–Nowcast (WRE–N) Forecasts Using a Spatial–Categorical Method

by John W Raby
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* July 2017	**2. REPORT TYPE** Technical Report	**3. DATES COVERED (From - To)** October 2016–July 2017

4. TITLE AND SUBTITLE

Verification of Weather Running Estimate–Nowcast (WRE–N) Forecasts Using a Spatial–Categorical Method

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

John W Raby

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

US Army Research Laboratory
Computational and Information Sciences Directorate (ATTN: RDRL-CIE-M)
White Sands Missile Range, NM 88002

8. PERFORMING ORGANIZATION REPORT NUMBER

ARL-TR-8064

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Spatial forecasts from Numerical Weather Prediction (NWP) models of meteorological variables to support US Army operations on the battlefield have become an integral part of the products available for the Staff Weather Officer to use in providing mission planning and execution forecasts. These forecasts are ingested by certain Army tactical decision aids (TDAs) and are fused with information on the operational weather thresholds, which impact the performance of Army systems and missions. Such a TDA generates spatial and temporal forecasts of these impacts for user-specified systems and/or missions. This report presents the results from applying a method to verify forecast fields of meteorological variables that have been filtered by the application of a threshold, similar to the method used by the TDA. A threshold applied to a continuous variable field becomes a categorical forecast for which there are traditional and nontraditional methods for verification. This study evaluates the ability of the NWP model to predict multiple categories of the spatial variable. Preliminary results suggest the skill of the model when predicting objects defined by lower thresholds is greater than the skill for objects defined by higher thresholds.

15. SUBJECT TERMS

forecast, verification, categorical forecast, weather impacts, thresholds, numerical weather prediction, observations, Model Evaluation Tools, MET Series-Analysis, My Weather Impacts Decision Aid, MyWIDA, tactical decision aid, TDA

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** John W Raby
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	UU	36	**19b. TELEPHONE NUMBER (Include area code)** 575-678-2004

Contents

List of Figures

List of Tables

Preface

This technical report relates to a previous work that explores the application of categorical and object-based verification methods to verify spatial forecasts produced by the Weather Running Estimate–Nowcast (WRE–N) of continuous meteorological variables that have been filtered by a single threshold. These methods use gridded forecasts and observations on a common grid, which enables the application a number of different spatial verification methods that reveal various aspects of model performance. This report describes the results obtained when the same categorical method, called "spatial categorical" in this report, was applied to the same data to determine the ability of the WRE–N to predict objects defined by multiple thresholds. Thus, portions of this report's content originated in ARL-TR-7751.[1]

[1] 1 Raby JW, Cai H. Verification of spatial forecasts of continuous meteorological variables using categorical and object-based methods. White Sands Missile Range (NM): Army Research Laboratory (US); 2016 Aug. Report No.: ARL-TR-7751.

Acknowledgments

The author offers his thanks to Mr Robert Dumais of the US Army Research Laboratory (ARL), who contributed data and information without which the study could not have been completed; Dr Huaqing Cai of ARL, whose suggestions strengthened and clarified the analysis and conclusions of the report; and Mr Jeff Johnson, whose suggestions improved and clarified the implications of the results of this study on Army tactical decision aids (TDAs) and specifically to the My Weather Impacts Decision Aid TDA.

Many thanks to Mr Martin Kufus of the ARL Technical Publishing office at White Sands Missile Range, New Mexico, for his many helpful suggestions and consistently high standard of editing.

Executive Summary

Spatial forecasts from Numerical Weather Prediction (NWP) models of tactically significant meteorological variables to support US Army operations on the battlefield have become an integral part of the products available for the Air Force Staff Weather Officer to use in providing mission planning and execution forecasts. These forecasts are ingested by certain Army tactical decision aids (TDAs). Such TDAs fuse information on the characteristic operational weather thresholds that potentially affect (impact) missions and performance of systems conducting the missions with the spatial forecast information from NWPs. The TDA generates spatial forecasts of these impacts for user-specified systems and/or missions and for the time period and location of interest.[1] This report presents the results obtained by applying a spatial–categorical method that can verify spatial forecast fields of meteorological variables that have been filtered by the application of a threshold or category the same way as that used by the TDA. In effect, a threshold applied to a continuous variable field becomes a categorical forecast for which there are traditional and nontraditional methods for verification. This study evaluates the ability of the NWP model to predict multiple categories of the spatial variable and compares the skill of the model for the different categories.

Traditional methods have been developed to verify the skill of NWP to predict categories of continuous meteorological variables. These methods apply the established theoretical framework for evaluating deterministic binary forecasts. This framework involves defining a binary event through the application of a category or threshold and evaluates the forecast skill by counting the numbers of times the event was forecast or not and observed or not in a contingency table. There are numerous statistics and skill scores that can be computed from the data collected by this method. For this study, the author obtained forecasts from the Army's Weather Running Estimate–Nowcast, which is an Advanced Research version of the Weather Research and Forecasting Model adapted for generating short-range nowcasts and gridded observations produced by the National Oceanographic and Atmospheric Administration's Global Systems Division using the Local Analysis and Prediction System. A tool developed by the National Center for Atmospheric Research called MET Series-Analysis was used to generate the skill scores and statistics at every grid point; then, generate graphical products that display the spatial distribution of the scores and statistics for each of 4 categories.

[1] Johnson J. Personal communication. White Sands Missile Range (NM): Army Research Laboratory (US); 2017 June 17.

Preliminary results suggest the skill of the model when predicting objects defined by lower thresholds is greater than the skill for objects defined by higher thresholds.

1. Introduction and Background

As computing technology has advanced, the weather-forecasting task, once the primary role of a human forecaster in theater, has shifted to computerized Numerical Weather Prediction (NWP) models. Scientists around the world have used the Weather Research and Forecasting model (WRF) extensively for many applications. In this study, the model used was the Advanced Research version of WRF (Skamarock et al. 2008) that we abbreviate as WRF–ARW. WRF–ARW includes Four-Dimensional Data Assimilation (FDDA) techniques that can be used to incorporate observations into the model so that forecast quality is improved (Stauffer and Seaman 1994; Deng et al. 2009). The US Army Research Laboratory (ARL) uses WRF–ARW as the core of its Weather Running Estimate–Nowcast (WRE–N) weather-forecasting model.

The Army requires high-resolution weather forecasting to model atmospheric features with wavelengths on the order of 5 km or less; that imposes a requirement for NWP to operate on a model grid spacing on the order of 1 km or less in the finest, or most resolved, domain to resolve weather phenomena of interest to the Soldier in theater. The atmospheric flows of interest to the Army include mountain/valley breezes, sea breezes, and other flows induced by differences in land-surface characteristics. High-resolution NWP forecasts need to be validated against observations before their outputs can be used effectively by My Weather Impacts Decision Aid (MyWIDA), an Army-developed decision aid used to determine atmospheric impacts on Army and Joint systems and operations (Brandt et al. 2013). Weather-forecast validation has always been of interest to the civilian and military weather-forecasting community; see, for example, the reviews by Ebert et al. (2013) and Casati et al. (2008) or the books by Jolliffe and Stephenson (2012) or Wilks (2011). The validation of the models, especially high-resolution NWP, has proven to be especially difficult when addressing small temporal and spatial scales (NRC 2010) that characterize NWP for use in Army applications. Furthermore, the verification of WRE–N spatial fields of continuous meteorological variables that have been filtered by the application of a threshold has not been accomplished.

The WRF model is maintained by the National Center for Atmospheric Research (NCAR), which has also developed a suite of Model Evaluation Tools (MET) (NCAR 2013) to evaluate WRF–ARW performance. MET was developed at NCAR through a grant from the US Air Force 557th Weather Wing (formerly the Air Force Weather Agency). NCAR is sponsored by the National Science Foundation. MET Series-Analysis performs spatial–categorical verification of gridded model output against observations that have been analyzed and placed on a grid matching that of the model.

ARL has employed MET Series-Analysis in a prior study, the results of which are presented by Raby and Cai (2016). They evaluated the applicability of a combination of a categorical and object-based technique for assessing the 1.75-km grid spacing WRE–N model to demonstrate the utility of combining traditional and nontraditional techniques for assessing the ability of the model to predict objects defined by application of a single threshold.

ARL's collaborations with the National Oceanic and Atmospheric Administration's (NOAA's) Global Systems Division (GSD) resulted in the generation of 1.75-km grids of observations of surface meteorological variables for the same domain as the WRE–N using the NOAA–GSD Local Analysis and Prediction System (LAPS).

The WRE–N was run with and without FDDA for 5 case-study days over a 1.75-km grid-spacing domain in Southern California over highly varied terrain and with a dense observational network that provided a robust data set of model output for analysis. Since results from a comparison of the verification skill scores for the FDDA runs with those run without the FDDA showed nearly identical scores (Raby and Cai 2016), only the model runs with FDDA were used for this study. The case-study days from February–March 2012 were picked to vary weather conditions from a strong synoptic forcing situation to a quiescent situation. (The weather conditions for each study day are described in Section 2.3.)

This study employs MET Series-Analysis to generate spatial–categorical-verification results for assessing the WRE–N at tactically significant grid spacings for a range of threshold values applied to forecasts of continuous meteorological variables. The motivation for presenting results at multiple thresholds came from a suggestion by a colleague who posed a question about the performance of the model at lower thresholds in view of lower skill when predicting objects defined by the highest threshold (Jameson 2016). The skill scores generated at a given threshold provide an assessment of the ability of the model to predict the object defined by the threshold similar to the way MyWIDA uses output from models such as WRE–N to provide spatial distributions of forecast weather impacts to Army missions and systems. By design and intent, Army systems and missions must be able to operate in all weather conditions, but there are rules that define marginal and unfavorable conditions in terms of numerous meteorological variables that are intended to serve as a general guide for decision-makers to consider before planning or executing an operation. For unfavorable impacts due to a single variable, MyWIDA typically uses a single threshold—"greater than or equal" (GE) or "less than or equal"—for a given variable based on the rules that define the unfavorable weather impacts on systems or missions. Unfavorable conditions are usually associated with the most extreme condition that adversely impacts the system or mission.

2. Domain and Model

The ARL WRE–N (Dumais et al. 2004; Dumais et al. 2013) has been designed as a convection-allowing application of the WRF–ARW model (Skamarock et al. 2008) with an observation-nudging FDDA option (Liu et al. 2005; Deng et al. 2009). For this investigation, the WRE–N was configured to run over a multinest set of domains to produce a fine inner mesh with 1.75-km grid spacing, and it leveraged an external global model for cold-start initial conditions and time-dependent lateral boundary conditions for the outermost nest. Table 1 describes the dimensions for the triple-nested domain. This global model for ARL development and testing has been the National Centers for Environmental Prediction's Global Forecast System (GFS) model (EMC 2003). The WRE–N is envisioned to be a rapid-update cycling application of WRF–ARW with FDDA and optimally could refresh itself at intervals up to hourly (dependent upon the observation network) (Dumais et al. 2012; Dumais and Reen 2013).

Table 1 WRE–N triple-nested domain dimensions in km

East–West dimension	North–South dimension	Grid spacing
1780	1780	15.75
761	761	5.25
506	506	1.75

For this study, the model runs had a base time of 1200 coordinated universal time (UTC) and produced output for each hour from 1200 UTC to 0600 UTC of the following day for a total of 19 hourly model outputs, which were produced for each of 5 days in February and March 2012. The modeling domains are depicted in Fig. 1.

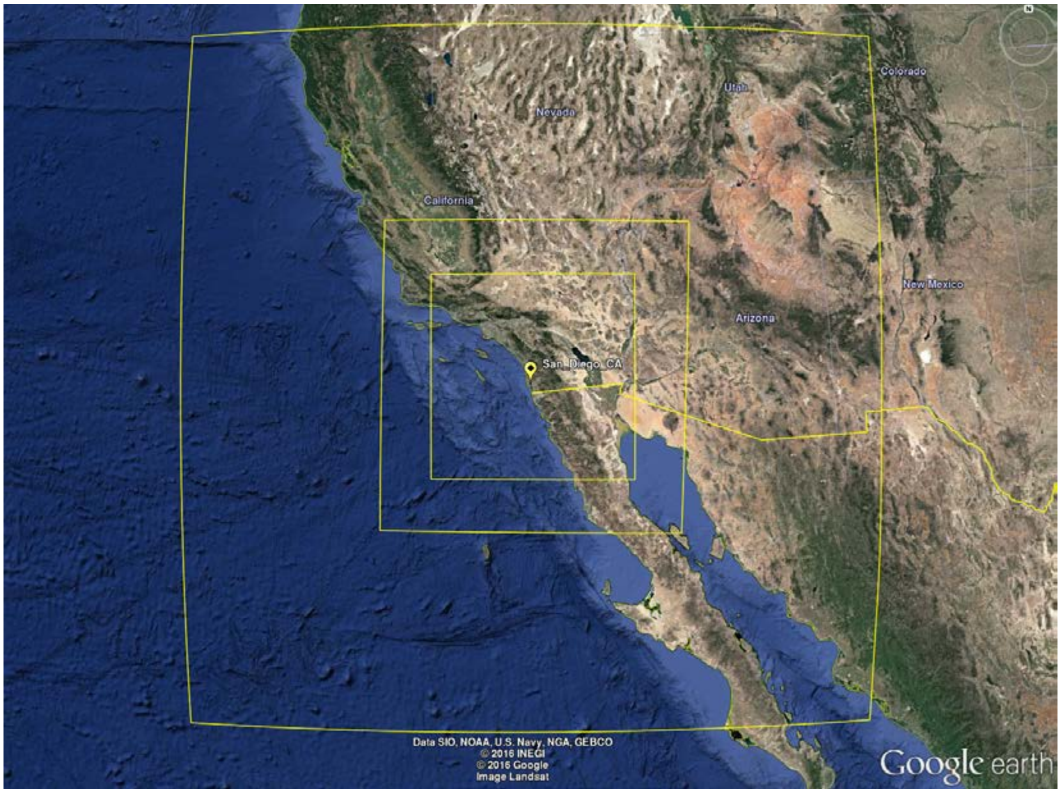

Fig. 1 Triple-nested model domains; domain center points are coincident and centered near San Diego, California (Google Earth 2016)

2.1 Observations for Assimilation

The initial conditions were constructed by starting with the GFS data as the first guess for an analysis using observations. Most observations were obtained from the Meteorological Assimilation Data Ingest System (MADIS) (NOAA 2016), except for the Tropospheric Airborne Meteorological Data Reporting (TAMDAR) (Daniels et al. 2016) observations, which were obtained from AirDat, LLC. The MADIS database included standard surface observations, mesonet* surface observations, maritime surface observations, wind-profiler measurements, rawinsonde soundings, and Aircraft Communications, Addressing, and Reporting System (ACARS) data. Use and reject lists were obtained from developers of the RTMA system (De Pondeca et al. 2011), and these were used to filter MADIS mesonet observations. This quality-assurance evaluation is especially important given the greater tendency of mesonet observations to be more poorly sited than other, more standard, surface observations.

The Obsgrid component of WRF was used for quality control of all observations. This included gross-error checks, comparison of observations to a background field

* A network of automated meteorological observation stations.

(here GFS), and comparison of observations to nearby observations. We modified Obsgrid to allow observations such as the TAMDAR and ACARS data to be more effectively compared against the GFS background field. The quality-controlled observations were output in hourly, “little_r” formatted text files for use as ground-truth data for model assessment. We employed observation nudging to the observations from these same sources for the preforecast period of 1200–1800 UTC (0- through 6-h lead times), followed by 1 h ramping down of the nudging from 1800 to 1900 UTC, during which no new observations are assimilated. The true, free forecast period thus begins at 1800 UTC because no observations after this time are assimilated.

2.2 Parameterizations

For the parameterization of turbulence in WRE–N, a modified version of the Mellor–Yamada–Janjić (MYJ) Planetary Boundary Layer (PBL) (Janjić 1994) scheme was used. This modification decreases the background turbulent kinetic energy and alters the diagnosis of the boundary-layer depth used for model output and data assimilation (Reen et al. 2014). The WRF single-moment, 5-class microphysics parameterization is used on all domains (Hong et al. 2004), while the Kain–Fritsch (Kain 2004) cumulus parameterization is used only on the 15.75-km outer domain. For radiation, the Rapid Radiative Transfer Model (RRTM) parameterization (Mlawer et al. 1997) is used for longwave radiation and the Dudhia (1989) scheme for shortwave radiation. The Noah land-surface model (Chen and Dudhia 2001a, 2001b) is used. Additional references and other details for these parameterization schemes are available from Skamarock et al. (2008). Table 2 lists the WRF configuration settings.

Table 2 WRE–N configuration

Configuration	
WRF–ARW V3.4.1	Yes
Obs-nudging FDDA	Yes
Multinest (15.75/5.25/1.75 km)	Yes
MADIS observations (FDDA)	Yes
TAMDAR observations (FDDA)	Yes
Ship/buoy observations (FDDA)	Yes
Filter obs (use/reject) (FDDA)	Yes
RUNWPSPLUS quality control (FDDA)	Yes
Obs-nudge rad 120,60,20	Yes
MYJ–PBL scheme (modified)	Yes
WRF,sgl-moment, 5-class microphysics	Yes
Option 8—microphysics	Yes
End FDDA 360 min	Yes
Kain–Fritsch Cum Param (outer domain)	Yes
RRTM long-wave rad (Mlawer)	Yes
Shortwave rad (Dudhia)	Yes
Noah land-surface model	Yes
Fix for nudge to low water vapor	Yes
Model Top 10hPa	Yes
Feedback on	Yes
Obs weighting function 4E-4	Yes
57 vertical levels	Yes
48-s time step	Yes

2.3 Case-Study Days

The case-study days were selected on the basis of the prevailing synoptic weather conditions over the nested domains. Table 3 provides a short description of these conditions.

Table 3 Synoptic conditions for the case-study days considered

Case	Dates (all 2012)	Description
1	February 07–08	Upper-level trough moved onshore, which led to widespread precipitation in the region.
2	February 09–10	Quiescent weather was in place with a 500-hPa ridge centered over central California at 1200 UTC.
3	February 16–17	An upper-level low located near the California–Arizona border with Mexico at 1200 UTC brought precipitation to that portion of the domain. This pattern moved south and east over the course of the day.
4	March 01–02	A weak shortwave trough resulted in precipitation in northern California at the beginning of the period that spread to Nevada, then moved southward and decreased in coverage.
5	March 05–06	Widespread high-level cloudiness due to weak upper-level low pressure but very limited precipitation.

2.4 Observations for Verification

The LAPS gridded observation data sets produced by NOAA–GSD consisted of 12 hourly Gridded Binary format, edition 2 (GRIB2) files of 2-m above-ground-level (AGL) temperature (TMP), relative humidity (RH), and dew-point temperature (DPT) and 10-m AGL U-component and V-component winds for the period of 1200–2300 UTC (forecast lead times 0 through 11) on each of the 5 cases. The output grid used by the LAPS was 289 × 289 with 1.75-km grid spacing.

3. Data Preparation Using MET

The model and observational data were preprocessed into the formats required by MET Series-Analysis. The WRE–N model output data were converted from native Network Common Data Form (NetCDF) files to hourly Gridded Binary format, edition 1 (GRIB) files by the WRF Unified Post Processor, which destaggers the data onto an Arakawa-A Grid containing 288 × 288 grid points. The hourly GRIB2 files on a 289 × 289 grid had to be remapped to the 288 × 288 grid to match that of the WRE–N grid. The NCAR "COPYGB" utility program was used to remap the observations and convert the files to GRIB (DTC 2016). The author used MET Series-Analysis to generate the grid-to-grid, categorical-error statistics for surface meteorological variables TMP and DPT in degrees Kelvin (K), RH (%), and wind speed in meters per second (WIND) for every grid point in the model domain to provide a way to see the spatial distribution of the errors. Series-Analysis computed the contingency-table statistics and skill scores for each forecast hour for 5 different thresholds (categories) at every grid point over all 12 forecast lead times and all 5 case-study days. The thresholds were specified using the FORTRAN convention of "GE" to indicate greater than or equal to the given threshold value and are shown in Table 4.

Table 4 Thresholds used in MET Series-Analysis

TMP (K)	DPT (K)	RH (%)	WIND (m/s)
270	262	25	2
275	267	40	5
280	272	55	8
285	277	70	11
290	282	85	14

MET Series-Analysis generates many categorical skill scores and contingency-table statistics. Of these, Table 5 lists those which were output initially.

Table 5 Initial Series-Analysis skill scores and contingency-table statistics

Score/statistic	Description
BASER	base rate
FMEAN	mean forecast value
PODY	hit rate
FAR	false-alarm ratio
FBIAS	frequency bias
CSI	Critical Success Index
GSS	Gilbert Skill Score
ACC	accuracy

For this study, the author reduced the analysis to consider only CSI and FBIAS for the variables of 2-m AGL TMP and RH and 10-m AGL WIND to accomplish a preliminary assessment of the accuracy of WRE–N output that was filtered by application of multiple thresholds. Table 6 shows the variables and thresholds used in the analysis. The Series-Analysis output NetCDF file was ingested into the Unidata Integrated Data Viewer, which was used to generate graphics displaying the spatial distribution of the CSI and FBIAS over the WRE–N domain (Murray et al. 2003).

Table 6 Analysis thresholds

TMP	RH	WIND
(K)	(%)	(m/s)
290	85	11
285	70	8
280	55	5
275	40	2

4. Analysis of MET Series-Analysis Results

The CSI and FBIAS are defined by a ratio of counts determined using a 2 × 2 contingency table. Table 7 shows the contingency table with notation consistent with the formulae for the scores and statistics as implemented in the MET (NCAR 2013).

Table 7 2 × 2 contingency table from the MET User's Guide 4.1 (NCAR 2013)

Forecast	Observation		Total
	o = 1 (e.g., "Yes")	o = 0 (e.g., "No")	
F = 1 (e.g., "Yes")	n11	n10	n1. = n11 + n10
F = 0 (e.g., "No")	n01	n00	n0. = n01 + n00
Total	n.1 = n11 + n01	n.0 = n10 + n00	T = n11 + n10 + n01 + n00

[a] 2 × 2 contingency table in terms of counts. The n_{ij} values in the table represent the counts in each forecast-observation category, where *i* represents the forecast and *j* represents the observations. The "." symbols in the Total cells represent sums across categories.

[b] The counts, n_{11}, n_{10}, n_{01}, and n_{00}, are sometimes called the "hits", "false alarms", "misses", and "correct rejections", respectively.

[c] By dividing the counts in the cells by the overall total, *T*, the joint proportions, p_{11}, p_{10}, p_{01}, and p_{00} can be computed. Note that $p_{11} + p_{10} + p_{01} + p_{00} = 1$. Similarly, if the counts are divided by the row (column) totals, conditional proportions, based on the forecasts (observations) can be computed.

The CSI score (Eq. 1) is computed as described in the MET User's Guide 4.1 (NCAR 2013):

$$\mathrm{CSI} = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}\,, \tag{1}$$

with CSI being the ratio of the number of times the event was correctly forecasted to occur to the number of times it was either forecasted or occurred. CSI ignores the "correct rejections" category (i.e., n_{00}).

The value of the CSI ranges between 0 and 1, with 1 being a perfect forecast and 0 being a forecast with no skill.

The FBIAS score is computed as described below in Eq. 2:

$$\mathrm{Bias} = \frac{n_{11} + n_{10}}{n_{11} + n_{01}} = \frac{n_{1.}}{n_{.1}} \tag{2}$$

with FBIAS defined as the ratio of the total number of forecasts of an event to the total number of observations of the event. A "good" value of frequency bias is close to 1; a value greater than 1 indicates the event was forecasted too frequently and a value less than 1 indicates the event was not forecasted frequently enough.

4.1 Compare CSI and FBIAS for the 4 Threshold Values

A display of the spatial distribution of the CSI for TMP for 4 different thresholds is shown in Fig. 2. The plot for TMP GE 290 shows the CSI score for the case with the highest threshold that was generated for the previous study by Raby and Cai (2016). Note the areas that are white in color do not have a CSI score due to nonoccurrences of the GE 290-K event. The other plots show how CSI changes from generally lower CSI scores to higher scores as the threshold value is lowered. Visually, this trend appears as a transition from cooler to warmer colors with dark orange indicating a perfect CSI score of 1. At 275 K, the CSI over most of the domain is near perfect with slightly lower scores over mountainous terrain and the Sea of Cortez. This trend matches the expected trend as described by Jolliffe and Stephenson (2012).

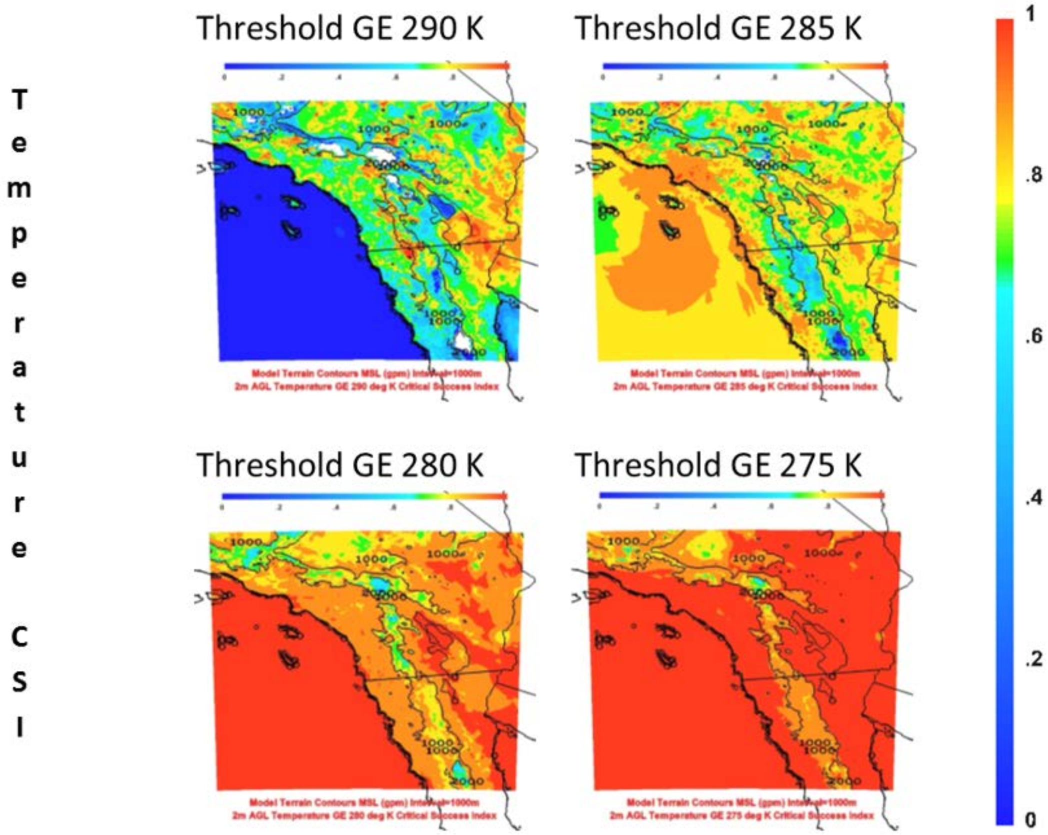
Temperature CSI
Threshold GE 290 K
Threshold GE 285 K
Threshold GE 280 K
Threshold GE 275 K
1000
2000
Model Terrain Contours MSL (gpm) Interval=1000m
2m AGL Temperature GE 290 deg K Critical Success Index
2m AGL Temperature GE 285 deg K Critical Success Index
2m AGL Temperature GE 280 deg K Critical Success Index
2m AGL Temperature GE 275 deg K Critical Success Index
1
.8
.6
.4
.2
0

Fig. 2 CSI for 2-m AGL TMP for 4 thresholds

A display of the spatial distribution of the FBIAS for TMP for 4 different thresholds is shown in Fig. 3.

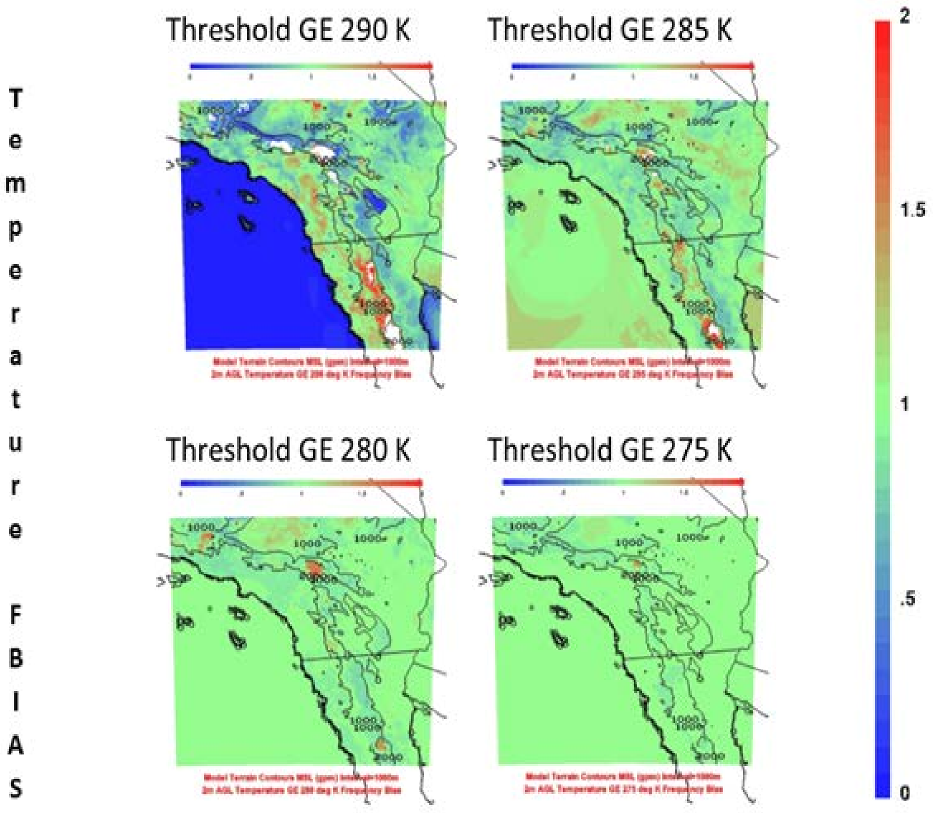
Temperature FBIAS
Threshold GE 290 K
Threshold GE 285 K
Threshold GE 280 K
Threshold GE 275 K
1000
2000
Model Terrain Contours MSL (gpm) Interval=1000m
2m AGL Temperature GE 290 deg K Frequency Bias
2m AGL Temperature GE 285 deg K Frequency Bias
2m AGL Temperature GE 280 deg K Frequency Bias
2m AGL Temperature GE 275 deg K Frequency Bias
2
1.5
1
.5
0

Fig. 3 FBIAS for 2-m AGL TMP for 4 thresholds

The plot for TMP GE 290 shows the FBIAS score for the case with the highest threshold that was generated for the previous study by Raby and Cai (2016). Note the areas that are white in color do not have a FBIAS score due to nonoccurrences of the GE 290-K event. The other plots show the same improving trend as that observed for CSI with decreasing bias as the threshold is lowered. Visually, this trend appears as a transition to the green color indicating an FBIAS score of 1 or no bias. Again, this trend agrees with the expected trend according to Jolliffe and Stephenson (2012). The WRE–N at the lowest threshold performs very well over almost the entire domain with almost no bias.

A display of the spatial distribution of the CSI for RH for 4 different thresholds is shown in Fig. 4.

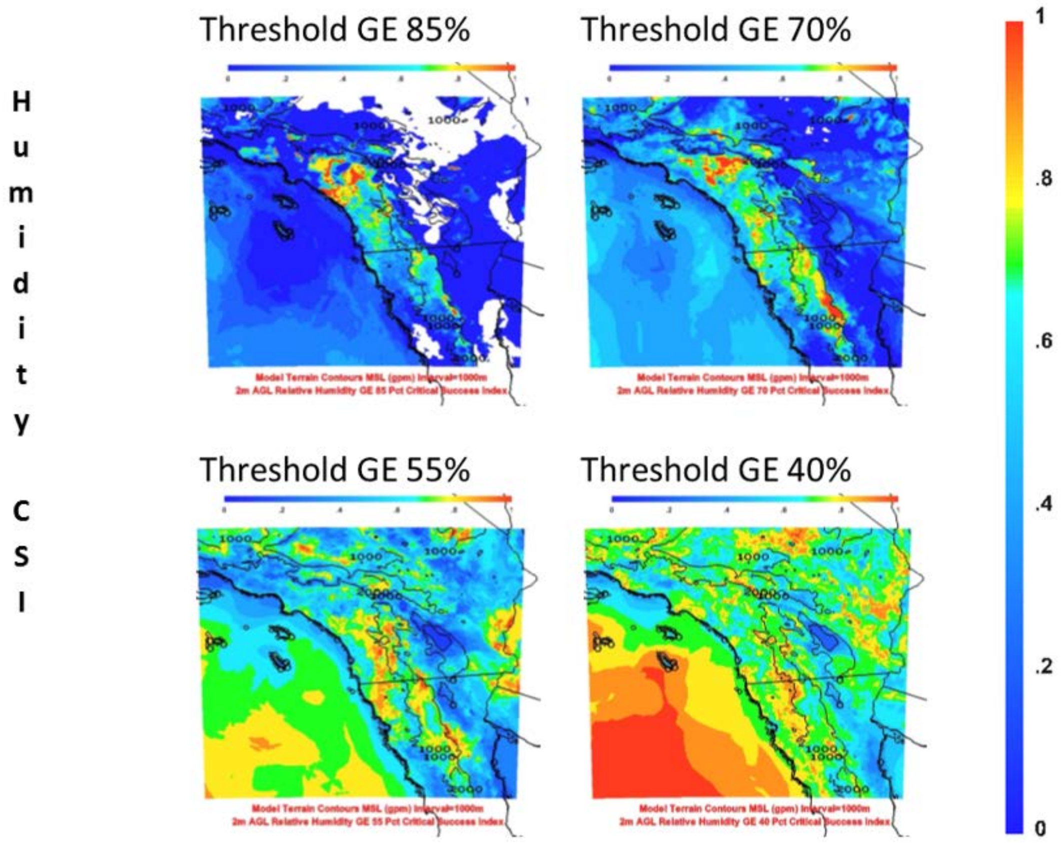
Humidity CSI
Threshold GE 85%
Threshold GE 70%
Threshold GE 55%
Threshold GE 40%
1000
2000
Model Terrain Contours MSL (gpm) Interval=1000m
2m AGL Relative Humidity GE 85 Pct Critical Success Index
2m AGL Relative Humidity GE 70 Pct Critical Success Index
2m AGL Relative Humidity GE 55 Pct Critical Success Index
2m AGL Relative Humidity GE 40 Pct Critical Success Index
1
.8
.6
.4
.2
0

Fig. 4 CSI for 2-m AGL RH for 4 thresholds

The plot for RH GE 85% shows the CSI score for the case with the highest threshold that was generated for the previous study by Raby and Cai (2016). Note the areas that are white in color do not have a CSI score due to nonoccurrences of the GE 85% event. The other plots show how CSI increases as the threshold value is lowered. Visually, this trend appears as a transition from cooler to warmer colors with dark orange indicating a perfect CSI score of 1. At 40%, the CSI over most areas of the domain has improved, especially over the ocean and to a lesser extent over land. This trend matches the expected trend as described by Jolliffe and Stephenson (2012).

A display of the spatial distribution of the FBIAS for RH for 4 different thresholds is shown in Fig. 5.

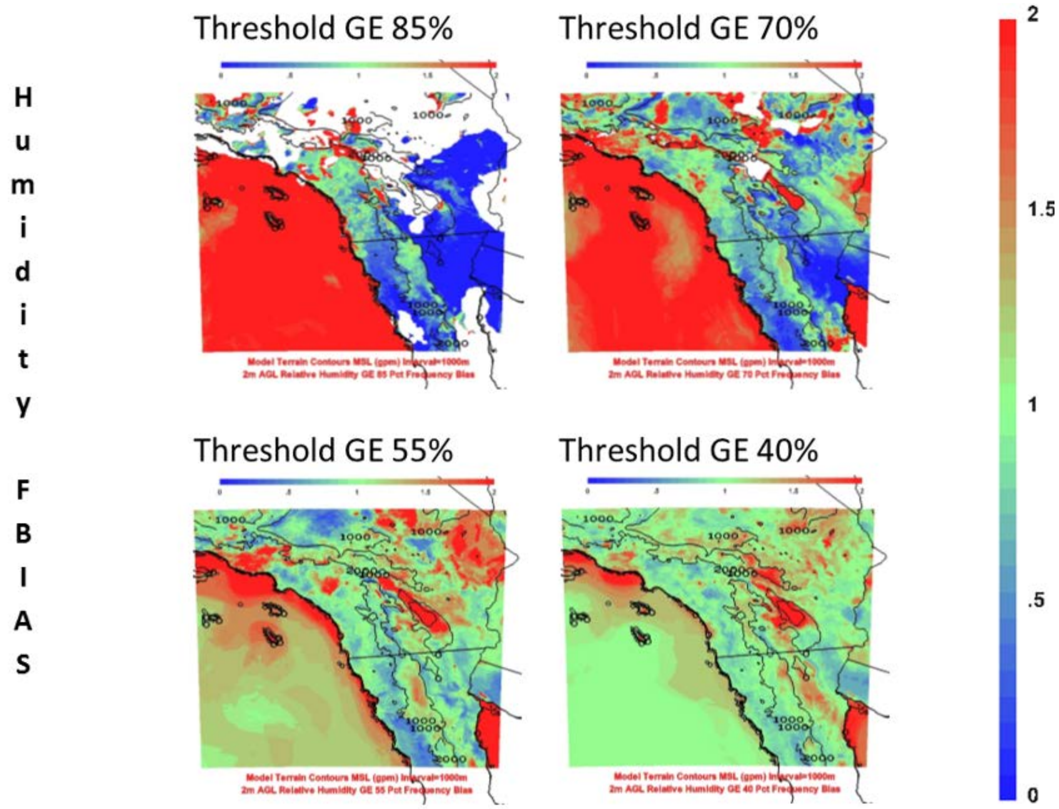
Humidity FBIAS
Threshold GE 85%
Threshold GE 70%
Threshold GE 55%
Threshold GE 40%
Model Terrain Contours MSL (gpm) Interval=1000m
2m AGL Relative Humidity GE 85 Pct Frequency Bias
2m AGL Relative Humidity GE 70 Pct Frequency Bias
2m AGL Relative Humidity GE 55 Pct Frequency Bias
2m AGL Relative Humidity GE 40 Pct Frequency Bias
1000
2000
2
1.5
1
.5
0

Fig. 5 FBIAS for 2-m AGL RH for 4 thresholds

The plot for RH GE 85% shows the FBIAS score for the case with the highest threshold that was generated for the previous study by Raby and Cai (2016). The other plots show the same improving trend as that observed for CSI with decreasing bias as the threshold is lowered. Note the areas that are white in color do not have an FBIAS score due to nonoccurrences of the GE 85% event and, to a lesser extent, the GE 70% event. Visually, the improving trend appears as a transition to the green color indicating an FBIAS score of 1 or no bias. Again, this trend agrees with the expected trend according to Jolliffe and Stephenson (2012). The WRE–N at the lowest threshold performs very well over a significant portion of the entire domain with almost no bias. The areas where there is an overforecasting bias appear to be those with lower elevation over land, the Salton Sea, the Sea of Cortez, and over the ocean in some parts of the coastal zone.

A display of the spatial distribution of the CSI for WIND for 4 different thresholds is shown in Fig. 6.

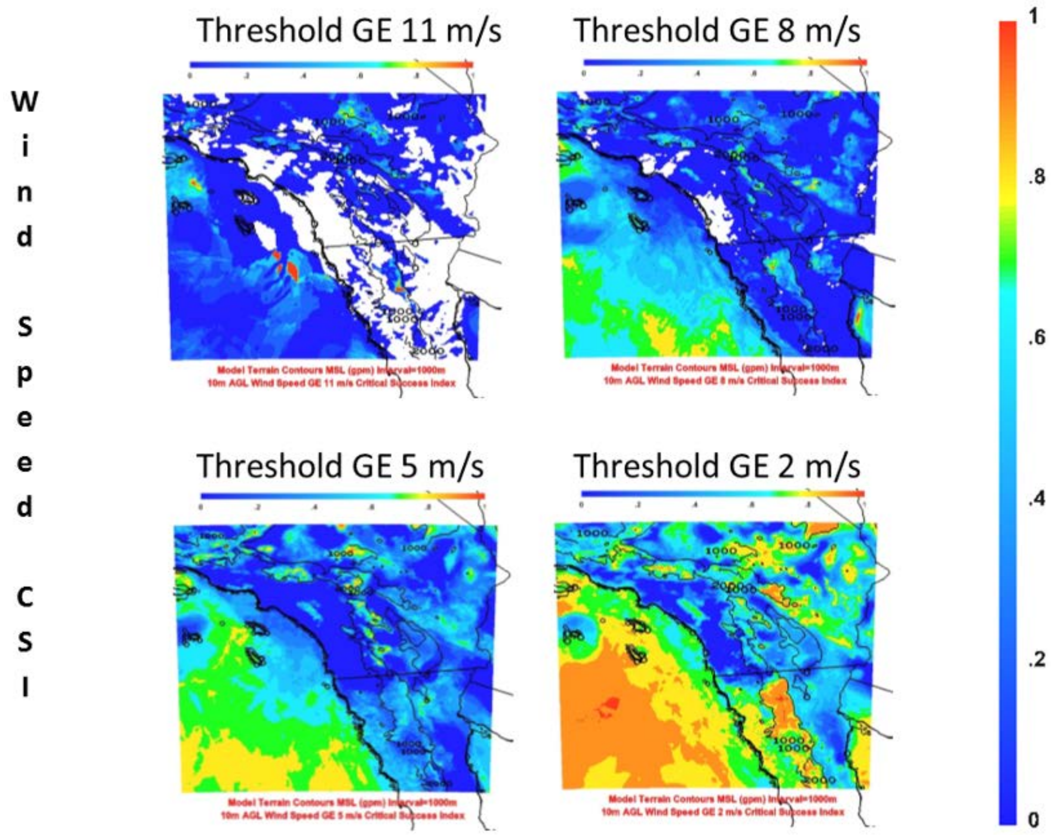
W i n d S p e e d C S I
Threshold GE 11 m/s
Threshold GE 8 m/s
Threshold GE 5 m/s
Threshold GE 2 m/s
Model Terrain Contours MSL (gpm) Interval=1000m
10m AGL Wind Speed GE 11 m/s Critical Success Index
10m AGL Wind Speed GE 8 m/s Critical Success Index
10m AGL Wind Speed GE 5 m/s Critical Success Index
10m AGL Wind Speed GE 2 m/s Critical Success Index
1000
2000
1
.8
.6
.4
.2
0

Fig. 6 CSI for 10-m AGL WIND for 4 thresholds

The plot for WIND GE 11 m/s shows the CSI score for the case with the highest threshold that was generated for the previous study by Raby and Cai (2016). The other plots show how CSI increases as the threshold value is lowered. Note the areas that are white in color do not have an FBIAS score due to nonoccurrences of the GE 11-m/s event and, to a lesser extent, the GE 8-m/s event. Visually, the improving trend appears as a transition from cooler to warmer colors with dark orange indicating a perfect CSI score of 1. At 2 m/s, the CSI over most areas of the domain has improved, especially over the ocean and the Sea of Cortez. This trend matches the expected trend as described by Jolliffe and Stephenson (2012).

A display of the spatial distribution of the FBIAS for WIND for 4 different thresholds is shown in Fig. 7.

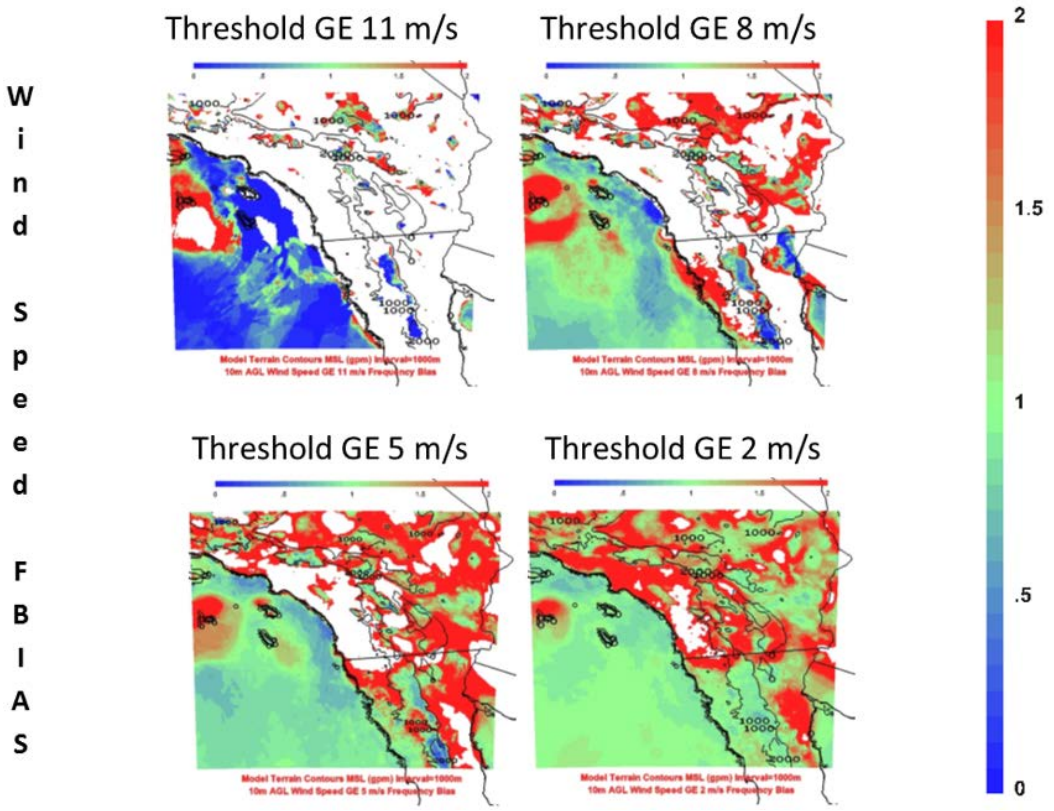
Wind Speed FBIAS
Threshold GE 11 m/s
Threshold GE 8 m/s
Threshold GE 5 m/s
Threshold GE 2 m/s
Model Terrain Contours MSL (gpm) Interval=1000m
10m AGL Wind Speed GE 11 m/s Frequency Bias
10m AGL Wind Speed GE 8 m/s Frequency Bias
10m AGL Wind Speed GE 5 m/s Frequency Bias
10m AGL Wind Speed GE 2 m/s Frequency Bias
1000
2000
2
1.5
1
.5
0

Fig. 7 FBIAS for 10-m AGL WIND for 4 thresholds

The plot for WIND GE 11 m/s shows the FBIAS score for the case with the highest threshold that was generated for the previous study by Raby and Cai (2016). The other plots show the same improving trend as that observed for CSI with decreasing bias as the threshold is lowered. Visually, this trend appears as a transition to the green color indicating an FBIAS score of 1 or no bias. Again, this trend agrees with the expected trend according to Jolliffe and Stephenson (2012). Note there are extensive areas of white indicating no occurrence of events defined by all 4 thresholds. Reducing the threshold resulted in a reduction of these nonevents. At the GE 2-m/s threshold, the remaining white areas are due to the nonoccurrence of observed winds that were GE 2 m/s, resulting in the FBIAS score being undefined. The WRE–N at the lowest threshold performs very well over a significant portion of that entire domain with almost no bias. The areas where there is an overforecasting bias appear to be mostly over land.

4.2 Summary of the Comparison of Scores for the 4 Threshold Values

The frequency of occurrence of forecast events determined by the application of thresholds to a continuous variable field changes spatially over the domain, affecting the CSI and FBIAS scores in a way that may give a misleading assessment of the model's ability to forecast objects. Analysis of these scores for a range of thresholds shows the WRE–N performs as expected with better scores achieved using lower threshold values.

When the thresholds are at the high end of the full range or, in some cases, the middle and lower segments of the range of the variable, there were areas where no events occurred, which limited the area where scores are calculated. Analysis of more categorical scores and contingency-table statistics—as well as assessment using object-based methods—is needed to overcome this limitation and improve assessments of the ability of the model to forecast objects defined using higher threshold values. Improved assessments of this aspect of model performance will lead to model improvements to enable better prediction of objects rendered using higher thresholds that will, in turn, translate into better MyWIDA unfavorable impact predictions.

The accuracy of the model judged from the scores varies considerably over the domain due to a combination of terrain characteristics and mesoscale variations in the air-mass characteristics. This is true of scores produced for all thresholds. Analysis of more scores and contingency-table statistics is needed to better relate them to terrain and air-mass characteristics. Use of a Geographic Information System (GIS) may be particularly useful for more in-depth error analysis based on domain partitioning. The implication of this variability suggests that weather impacts on Army systems and missions vary considerably in space.

The accuracy of the model at higher thresholds, judging from these results, is not as good as that using lower thresholds. The implication of this apparent lack of skill at higher thresholds is the prediction of unfavorable weather impacts generated by the MyWIDA TDA may not be as accurate as desired. However, for marginal weather impacts, which are associated with somewhat lower threshold values, the skill of the model may be better based on these results. Thus, it is important to conduct studies that use the actual system and mission thresholds to more accurately assess the ability of the model to predict objects that are meaningful to the Army. That said, use of actual thresholds will significantly reduce the number of locations and time periods for which the atmospheric conditions can provide data sets with the range of variable values that encompass actual thresholds. The impact of these 2 situations—each at odds with the other—has to be judged with the understanding that meaningful conclusions about model performance can only come from the analysis of large numbers of cases. So, there is a tradeoff between analysis of 1) data sets for fewer cases where tactically significant thresholds can be applied and 2) the more numerous data sets that were developed using thresholds defined by using the actual ranges of the variables present over the domain. The former presents challenges due to lack of statistically significant numbers of cases; the latter presents a challenge of limited application for assessment of the ability of models to forecast objects using mission- and system-specific thresholds.

5. Conclusion and Final Comments

The author found that the CSI and FBIAS skill scores produced using a spatial–categorical-verification method with multiple threshold values for each of the studied variables improve with decreasing threshold value. The amount of improvement was not the same over the entire domain, however. The study found that the frequency of occurrence of forecast events determined by the application of a high threshold value to a continuous variable field varies over the domain and affects the CSI and FBIAS scores in a way that may give a misleading assessment of the model's ability to forecast objects. Thresholds that define objects at the high end and, to a lesser extent, the mid- and lower portions of the range of a variable will produce scores over a subset of the domain because in some parts of the domain there were no event occurrences. This restricts the scoring to those areas where events occurred. As the threshold decreases, the numbers of nonevents decreases, allowing scores to be calculated over more of the domain. To more accurately assess the ability of the model to predict objects defined by high thresholds, studies are needed that use additional scores and statistics that are possible with the spatial–categorical method. Further, object-based methods provide additional information about the ability to predict objects. Raby and Cai (2016) recommended a more comprehensive approach combining several traditional and nontraditional methods for assessing the ability of the model to predict objects defined by thresholds; these numerous scores and statistics, when analyzed together, may reveal more information about model performance.

Another difficulty that arises when using high threshold values was discussed by the author (Raby 2016). The CSI and FBIAS scores presented in this report were reviewed by Cai (2016), who attributed the lack of skill at high thresholds to possibly the reduced size of objects that are defined by the high threshold, which leads to increases in model displacement errors. Raby (2016) presented results from object analysis at multiple thresholds showing the objects defined at low thresholds were larger than objects defined at high thresholds. For a given model displacement error, the resulting CSI scores indicate lower skill when the objects are small and indicate higher skill when the objects are larger. To illustrate this difference in scores, Fig. 8 depicts large and small objects and a given displacement error.

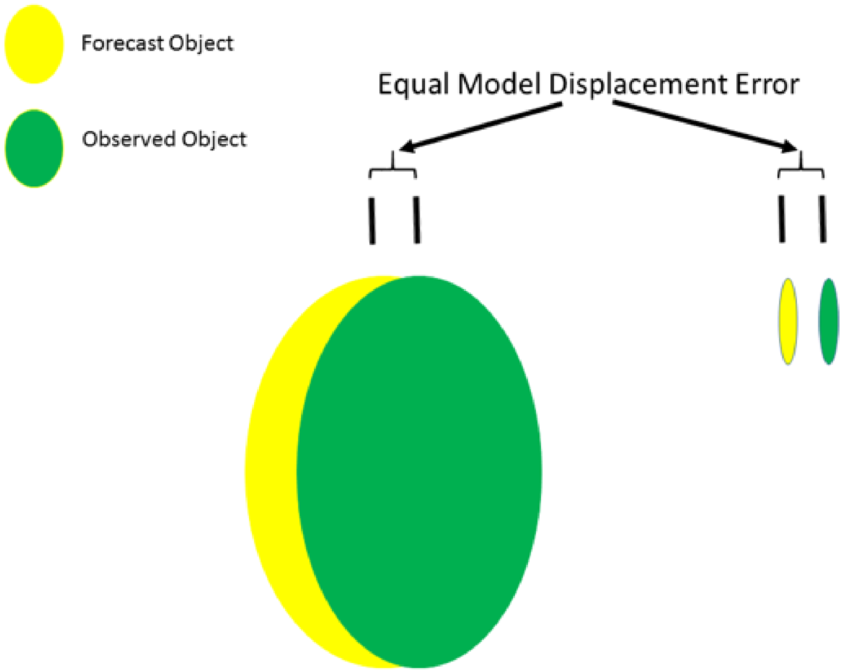

Fig. 8 Object-displacement error for large and small objects

The CSI is calculated from contingency-table statistics and is the ratio of the number of hits to the sum of the hits, false alarms, and misses. Figure 8 shows there is considerable agreement for the large objects despite the horizontal (east–west) displacement error. For the small objects, there is no agreement from the same displacement and there is the potential for more misses and less hits, especially if there are numerous small objects. The displacement error of small objects results in a significant decrease in the number of hits and increases in the number of misses, which serves to lower the CSI. By comparison, the same displacement error of large objects still results in a significant number of hits and thus decreases the number of misses, which serves to raise the CSI.

To further improve assessments of the predictability of objects, Raby and Cai (2016) recommended a more rigorous approach that requires the generation of larger data sets of forecast output and gridded observations so that statistically significant results can be obtained. This will be important when verifying the modeled objects defined at higher thresholds, particularly when WRE–N model output is used to predict the more critical unfavorable-weather impacts on Army systems and missions using MyWIDA.

Finally, to analyze and understand the complexity of the spatial variability of the scores revealed by this study and the previous study (Raby and Cai 2016), a GIS—which the atmospheric sciences have not extensively used—should be exploited for its ability to contextualize and analyze geospatial information such as terrain type/slope, land-use effects, and other spatial and temporal variables as explanatory metrics in model assessments (Smith et al. 2015, 2016a, 2016b). This technique has considerable promise of becoming an important new tool to augment other traditional and nontraditional tools for a comprehensive approach to model verification.

6. References

Brandt J, Dawson L, Johnson J, Kirby S, Marlin D, Sauter D, Shirkey R, Swanson J, Szymber R, Zeng S. Second generation weather impacts decision aid applications and web services overview. White Sands Missile Range (NM): Army Research Laboratory (US); 2013 July. Report No.: ARL-TR-6525.

Cai H. Personal communication. White Sands Missile Range (NM): Army Research Laboratory (US); 2016 Mar 15.

Casati B, Wilson LJ, Stephenson DB, Nurmi P, Ghelli A, Pocernich M, Damrath U, Ebert EE, Brown BG, Mason S. Forecast verification: current status and future directions. Meteo App. 2008;15(1):3–18.

Chen F, Dudhia J. Coupling an advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system. Part II: preliminary model validation. Mon Wea Rev. 2001a;129:587–604.

Chen F, Dudhia J. Coupling an advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system. Part I: model implementation and sensitivity. Mon Wea Rev. 2001b;129:569–585.

Daniels TS, Moninger WR, Mamrosh RD. Tropospheric airborne meteorological data reporting (TAMDAR) overview. Preprints, 10th Symposium on Integrated Observing and Assimilation Systems for Atmosphere, Oceans, and Land Surface; 2016 Sep 1; Atlanta (GA): American Meteorological Society [accessed 2016 Aug 2]. http://ams.confex.com/ams/pdfpapers/104773.pdf.

De Pondeca MSFV, Manikin GS, DiMego G, Benjamin SG, Parrish DF, Purser RJ, Wu WS, Horel JD, Myrick DT, Lin Y, et al. The real-time mesoscale analysis at NOAA’s National Centers for Environmental Prediction: current status and development. Wea Forec. 2011;26:593–612.

Deng A, Stauffer D, Gaudet B, Dudhia J, Hacker J, Bruyere C, Wu W, Vandenberghe F, Liu Y, Bourgeois A. Update on the WRF-ARW end-to-end multi-scale FDDA system. Paper presented at: 10th WRF Users' Workshop, National Center for Atmospheric Research, 2009 June 23–26; Boulder (CO).

[DTC] Developmental Testbed Center. MET online tutorial for METv3.0: COPYGB functionality. Boulder (CO): National Oceanic and Atmospheric Administration; [accessed 2016 July 27]. http://www.dtcenter.org/met/users /support/online_tutorial/METv3.0/copygb/index.php.

Dudhia J. Numerical study of convection observed during the Winter Monsoon Experiment using a mesoscale two-dimensional model. J Atmos Sci. 1989;46:3077–3107.

Dumais R, Kirby S, Flanigan R. Implementation of the WRF four-dimensional data assimilation method of observation nudging for use as an ARL Weather Running Estimate-Nowcast. White Sands Missile Range (NM): Army Research Laboratory (US); 2013 June. Report No.: ARL-TR-6485.

Dumais RE, Reen BP. Data assimilation techniques for rapidly relocatable weather research and forecasting modeling. White Sands Missile Range (NM): Army Research Laboratory (US); 2013 June. Report No.: Report ARL-TN-0546.

Dumais RE, Raby JW, Wang Y, Raby YR, Knapp D. Performance assessment of the three-dimensional wind field Weather Running Estimate-Nowcast and the three-dimensional wind field Air Force Weather Agency weather research and forecasting wind forecasts. White Sands Missile Range (NM): Army Research Laboratory (US); 2012 Dec. Report No.: ARL-TN-0514.

Dumais RE Jr, Henmi T, Passner J, Jameson T, Haines P, Knapp D. A mesoscale modeling system developed for the US Army. White Sands Missile Range (NM): Army Research Laboratory (US); 2004 Apr. Report No.: ARL-TR-3183.

Ebert E, Wilson L, Weigel A, Mittermaier M, Nurmi P, Gill P, Gober M, Joslyn S, Brown B, Fowler T, et al. Progress and challenges in forecast verification. Meteo App. 2013;20(2):130–139.

[EMC] Environmental Modeling Center. The GFS atmospheric model. Washington (DC): National Weather Service–National Centers for Environmental Prediction; 2003 Nov. NCEP Office Note No.: 442.

Google Earth. Mountain View (CA); 2016 [accessed 2016 Aug 24]. http://maps.google.com/help/terms_maps.html.

Hong SY, Dudhia J, Chen SH. A revised approach to ice microphysical processes for the bulk parameterization of clouds and precipitation. Mon Wea Rev. 2004;132:103–120.

Jameson T. Personal communication. White Sands Missile Range (NM): Army Research Laboratory (US); 2016 Sep 14.

Janjic ZI. The step-mountain eta coordinate model: further developments of the convection, viscous sublayer, and turbulence closure schemes. Mon Wea Rev. 1994;122:927–945.

Johnson J. Personal communication. White Sands Missile Range (NM): Army Research Laboratory (US); 2017 June 17.

Jolliffe IT, Stephenson DB. Forecast verification: a practitioner’s guide in atmospheric science. 2nd ed. Hoboken (NJ): John Wiley and Sons; 2012.

Kain JS. The Kain-Fritsch convective parameterization: an update. J App Meteo. 2004;43:170–181.

Liu Y, Bourgeois A, Warner T, Swerdlin S, Hacker J. Implementation of observation-nudging based FDDA into WRF for supporting ATEC test operations. Paper presented at: 6th WRF/15th MM5 Users’ Workshop, National Center for Atmospheric Research; 2005 June 27–30; Boulder, CO.

Mlawer EJ, Taubman SJ, Brown PD, Iacono MJ, Clough SA. Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. J Geoph Res Atmos. 1997;102:16663–16682.

Murray D, McWhirter J, Wier S, Emmerson S. The integrated data viewer—a web-enabled application for scientific analysis and visualization. Paper presented at: 19th International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology; 2003.

[NCAR] National Center for Atmospheric Research. Model evaluation tools version 4.1 (METv4.1), user’s guide 4.1. Boulder (CO); 2013 May.

[NOAA] Meteorological assimilation data ingest system (MADIS). College Park (MD): National Oceanic and Atmospheric Administration [accessed 2016 July 27]. http://madis.noaa.gov.

[NRC] National Research Council. When weather matters: science and service to meet critical societal needs. Washington (DC): The National Academies Press; 2010.

Raby JW. Application of a fuzzy verification technique for assessment of the Weather Running Estimate–Nowcast (WRE–N) model. White Sands Missile Range (NM): Army Research Laboratory (US); 2016 Oct. Report No.: ARL-TR-7849.

Raby JW, Cai H. Verification of spatial forecasts of continuous meteorological variables using categorical and object-based methods. White Sands Missile Range (NM): Army Research Laboratory (US); 2016 Aug. Report No.: ARL-TR-7751.

Reen BP, Schmehl KJ, Young GS, Lee JA, Haupt SE, Stauffer DR. Uncertainty in contaminant concentration fields resulting from atmospheric boundary layer depth uncertainty. J App Meteo Clim. 2014;53:2610–2626.

Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker DM, Duda MG, Huang XY, Wang W, Powers JG. A description of the advanced research WRF version 3. Boulder (CO): National Center for Atmospheric Research (US); 2008 June. NCAR Technical Note No.: TN-475-STR.

Smith JA, Foley TA, Raby JW, Reen B. Investigating surface bias errors in the Weather Research and Forecasting (WRF) model using a Geographic Information System (GIS). White Sands Missile Range (NM): Army Research Laboratory (US); 2015 Feb. Report No.: ARL-TR-7212.

Smith JA, Foley TA, Raby JW, Reen BP, Penc RS. Case study applying GIS tools to verifying forecasts over a domain. Paper presented at: 96th Annual Meeting of the American Meteorological Society, 23rd Conference on Probability and Statistics in the Atmospheric Sciences; 2016a Jan; New Orleans, LA.

Smith JA, Raby JW, Foley TA, Reen BP, Penc RS. Case study applying GIS tools to verifying forecasts over a mountainous domain. Paper presented at: 17th Mountain Meteorology Conference, American Meteorological Society; 2016b; Burlington, VT.

Stauffer DR, Seaman NL. Multiscale four-dimensional data assimilation. J App Meteo. 1994;33:416–434.

Wilks DS. Statistical methods in the atmospheric sciences. 3rd ed. Oxford (England): Academic Press; 2011.

List of Symbols, Abbreviations, and Acronyms

ACARS	Aircraft Communications, Addressing, and Reporting System
AGL	above ground level
ARL	US Army Research Laboratory
ARW	Advanced Research Weather Research and Forecasting model
CSI	Critical Success Index
DPT	dew-point temperature
FBIAS	Frequency Bias
FDDA	Four-Dimensional Data Assimilation
GE	greater than or equal to
GFS	Global Forecast System
GIS	Geographic Information System
GRIB	Gridded Binary format, edition 1
GRIB2	Gridded Binary format, edition 2
GSD	Global Systems Division
LAPS	Local Analysis and Prediction System
MADIS	Meteorological Assimilation Data Ingest System
MET	Model Evaluation Tools
MYJ	Mellor–Yamada–Janjic
MyWIDA	My Weather Impacts Decision Aid
NCAR	National Center for Atmospheric Research
NetCDF	Network Common Data Form
NOAA	National Oceanic and Atmospheric Administration
NWP	Numerical Weather Prediction
PBL	Planetary Boundary Layer
RH	relative humidity
RRTM	Rapid Radiative Transfer Model

RTMA	Real-Time Mesoscale Analysis
TAMDAR	Tropospheric Airborne Meteorological Data Reporting
TDA	Tactical Decision Aid
TMP	temperature
UTC	Coordinated Universal Time
WIND	wind speed
WRE–N	Weather Running Estimate–Nowcast
WRF	Weather Research and Forecasting
WRF–ARW	Weather Research and Forecasting, Advanced Research WRF

1 DEFENSE TECHNICAL
(PDF) INFORMATION CTR
DTIC OCA

2 DIRECTOR
(PDF) US ARMY RESEARCH LAB
RDRL CIO L
IMAL HRA MAIL & RECORDS
MGMT

1 GOVT PRINTG OFC
(PDF) A MALHOTRA

12 US ARMY RSRCH LAB
(PDF) RDRL CIE
P CLARK
T JAMESON
RDRL CIE M
J RABY
H CAI
B MACCALL
J SMITH
J PASSNER
R PENC
R DUMAIS
B REEN
RDRL CIE D
D KNAPP
J JOHNSON

1 US NAVY RSRCH LAB
(PDF) J MCLAY

1 US AIR FORCE
(PDF) R CRAIG

1 DCGS-A WX SVCS
(PDF) J CARROLL

3 UCAR
(PDF) T FOWLER
J H GOTWAY
B BROWN

1 USAICOE
(PDF) J STALEY